Missing Network Data A Comparison of Different Imputation Methods

Krause, Robert W.; Huisman, Mark; Steglich, Christian; Snijders, Tom A. B.

DOI:
10.1109/ASONAM.2018.8508716

Publication date:
2018

2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Missing Network Data
A Comparison of Different Imputation Methods

Robert W Krause
Dept. of Sociology
University of Groningen
Groningen, The Netherlands
r.w.krause@rug.nl

Mark Huisman
Dept. of Sociology
University of Groningen
Groningen, The Netherlands
j.m.e.huisman@rug.nl

Christian Steglich
Depts. of Sociology
University of Groningen
Groningen, The Netherlands
Institute for Analytical Sociology
Linköping, Sweden
c.e.g.steglich@rug.nl

Tom AB Snijders
Depts. of Sociology
University of Groningen
Groningen, The Netherlands
Nuffield College
Oxford, United Kingdom
t.a.b.snijders@rug.nl

***Abstract*—This paper compares several imputation methods for missing data in network analysis on a diverse set of simulated networks under several missing data mechanisms. Previous work has highlighted the biases in descriptive statistics of networks introduced by missing data. The results of the current study indicate that the default methods (analysis of available cases and null-tie imputation) do not perform well with moderate or large amounts of missing data. The results further indicate that multiple imputation using sophisticated imputation models based on exponential random graph models (ERGMs) lead to acceptable biases even under large amounts of missing data.**

***Index Terms*—missing data, social networks, exponential random graph model, Bayesian ERGM, multiple imputation**

I. Introduction

Empirical network studies of social relations and their structure is especially affected by missing data. First, missing data is more likely to occur in network data collection, because network questionnaires are complex and often touch upon sensitive topics. Second, the refusal of one member of the network to participate will automatically lead to missing data for all members of the network. When participants provide information about their outgoing links, they also provide information about the incoming links of other members of the network. Networks are therefore affected much more strongly by missing data than non-network data. The effects of missing data on network structure and analysis and the investigation into treatment procedures are an ongoing field of research [1], [2], [3]. In this study, we compare various techniques in their ability to capture key network descriptives under missing data. The methods (deletion, null-imputation and multiple imputation using Bayesian ERGMs) compete on a diverse set of simulated networks.

The paper is organized as follows. In Section II, we detail the topic of network analysis and the family of exponential random graph models. In Section III, we describe the non-response problem and its specifics for missing data in networks. We continue with a description of the data generation in Section IV. Section V presents the treatment methods used in this study. In Section VI, we present the results and we close the paper with a discussion of the findings and according recommendations.

IEEE/ACM ASONAM 2018, August 28-31, 2018, Barcelona, Spain
978-1-5386-6051-5/18/$31.00

II. Network Analysis

Network analysis is the study of sets of nodes and their links. These nodes can range from people to companies to countries, and the links can represent any form of relation from trade deals to interpersonal friendships. In the social sciences, the most common model family used to analyze cross-sectional social networks are exponential random graph models (ERGMs) [4], [5].

A. ERGMs and BERGMs

In short, ERGMs are probability models for networks where the probabilities depend on the frequency of occurrence of substructures in the network such as subgraph counts, or other statistics. Network structures are highly dependent upon each other, therefore testing hypotheses about structural properties of a network (e.g., girls are more likely to form cliques than boys) require to also model other network properties (e.g., the general tendency to form friendships). A sophisticated approach is needed because the dependencies between nodes and ties need to be taken into account. Networks are expressed as the random $n \times n$ adjacency matrix X with $X_{ij} = 1$ when there is tie from node i to node j and $X_{ij} = 0$ when there is no tie. Edges connecting nodes to themselves are not allowed ($X_{ii} = 0$). The networks can be directed or undirected (in that case $X_{ij} = X_{ji}$). Let $\mathbf{X}$ denote the set of all possible networks on n nodes and let x be a realization of X. In Bayesian ERGMs (BERGMs), as introduced by Caimo and Friel [6], the posterior conditional probability is given by

$$Pr(\theta|x) = \frac{\exp\left[\theta^T s(x)\right]}{z(\theta)} \frac{p(\theta)}{p(x)}, \tag{1}$$

with θ being a vector of model parameters, $s(x)$ a vector of corresponding sufficient statistics (e.g., number of edges or number of reciprocated ties), $z(\theta)$ the normalizing constant, $p(\theta)$ the prior distribution of the parameters and $p(x)$ is the marginal probability. See Lusher et al. for an introduction to ERGMs [4].

III. Missing Data

A. Missing Data Mechanisms

For an appropriate treatment of missing data in statistical modeling, it is important to consider the probability distribution of the missingness. Rubin defined three types of mechanisms for this probability distribution [7], which can be translated to the network data context [8]. First, data are missing completely at random (MCAR) if the probability of it to be missing is independent from any observed variable and also independent of the missing value itself. Second, data are called missing at random (MAR) if the probability of being missing is independent of the missing value but is dependent on other observed variables. For non-network data, treatment methods have been developed which yield unbiased estimates under these two mechanisms. The third mechanism is data missing not at random (MNAR). Data are MNAR if the probability of being missing is related to the missing value itself. This study will incorporate examples of all three missing data mechanisms.

B. Missing Data Types

While missing data mechanisms describe the probability distribution of the missing data, missing data types describe how the missingness is spread over the data set. In cross-sectional network research, two types of missing data can be distinguished: actor non-response and tie non-response [8]. Actor non-response occurs if all outgoing ties of an actor are missing. In tie non-response only some, but not all ties of an actor are missing. This study will focus only on actor non-response. However, the findings should also generalize to tie-nonresponse.

C. Effects of Missing Network Data

The effects of missing data on descriptive network statistics depend on the amount of missing data, on the network structure, on the descriptive statistic in question and how the missing data are treated. Some combinations of statistic and network are more robust to missingness than others. Larger and more centralized networks are usually more robust against missing data [9], and measures based on in-degrees are found to be generally reliable [3], [9], [10]. A notable difference between network and non-network data can be seen under the MCAR mechanism. While sample estimates of means, variances, and model parameters are usually unbiased for non-network data under MCAR, the same does not apply to network data. There are considerable biases found, even if data is missing completely at random [8], [9], [11].

D. Missing Data Treatments

Researchers have several options for handling missing data in networks. These options can broadly be separated into three categories: deletion, likelihood-based estimation and imputation (for a general overview of missing data handling see Schafer and Graham [12]). Deletion methods reduce the network to a fully observed subsample (listwise deletion of actors [8]) or ignore the missing data for some, but not all statistical calculations (pairwise deletion). Although deletion methods are commonly used and the default for most statistical programs, they do not perform well in most situations as they discard too much information [8], [11], [13].

Likelihood-based methods estimate the model parameters from the marginal distribution of the observed data. Under MAR this will lead to approximately unbiased estimates in larger samples, given that the used model is well fitting [12]. Likelihood-based estimation methods are available for different families of network models [5], [14], [15]. However, these methods are by definition model-based, and thus cannot aid the calculation of descriptive statistics or estimation of other models (e.g., blockmodels).

Imputation models replace the missing values with plausible guesses (for an overview of imputation methods for network data see Huisman and Krause [1]). Stochastic imputation methods use draws from probability distributions to replace missing values. These methods can be used for multiple imputation, where missing values are imputed multiple times based on a conditional probability model. This leads to a set of imputed data sets, which are analyzed separately leading to a distribution of model parameters. This allows for incorporating the uncertainty about the missing data imputation when estimating standard errors (for an introduction to multiple imputation see van Buuren [16]).

IV. Network Data

Before going into the missing data treatments compared in this study, we describe the (simulation of the) network data in more detail.

To be able to compare the performance of missing data treatment techniques for different networks, missing data mechanisms, and missing data rates, we simulated network data. Although results obtained from simulated data are harder to extrapolate to real, empirical data, they have several advantages over real world networks in the study of missing data.

First, the boundaries of simulated networks are clearly defined. While for data collection boundaries of empirical networks are generally well defined, actors usually also have links to actors outside the boundary. It is nearly impossible to collect the complete true neighborhood of all actors in a study. Although these effects are often marginal, simulations ensure that they are zero for all networks.

Second, all covariates are known. In empirical research there will always be a variable that has not been measured. Although often unrelated to the study, the complexities of the real world can be confounding factors and could lead to variations in the performance of missing data treatments across networks.

Third, we have control over the data generating process. This gives us experimental control over the network compositions in this study, thus allowing us to investigate the performance of the treatment methods under experimentally varying conditions. Further, it allows us to use the data generating model for imputation and allows us to investigate the performance of misspecified imputation models.

Fourth, using simulated networks ensures that there is no missing data in the complete observed network. Empirical network studies are likely to encounter missing data. Although it is vital to study empirical patterns of missing data in networks, they are a hindrance in evaluating missing data handling techniques and may even bias analyses. Knowing the true complete data allows the researcher to evaluate how well the treatment method performs and gives complete control over the missing data type and mechanism. In short, simulating the networks ensures that we can test the missing data techniques under optimal conditions.

A. Network Simulation

Networks were simulated using the ergm package in R [17], [18] including parameters for reciprocity, homophily, GWESP (geometrically weighted edgewise shared partners [19]) and GWDSP (geometrically weighted dyadwise shared partners [19]) while keeping the number of ties fixed. The networks differ in size (30 vs. 80 nodes), density (average degree 3 vs. 6), reciprocity (30% vs. 50% reciprocated ties) and homophily on a binary nodal covariate with half the group having the value 0 and the other half having the value 1 (50% vs. 70% homophilous ties). All networks have 30% transitive ties. This leads to 16 different configurations in total. For each configuration, ten complete networks were simulated, leading to 160 networks in total. The simulated networks were allowed to differ at maximum by 2.5% on any of the descriptive statistics. These configurations were selected such that the resulting simulated networks are similar in their structure to social networks that are often observed in small groups (e.g., classrooms).

B. Missing Data Creation

Missing data were created using six different mechanisms and five different missing data rates (10-50%). All missing data were generated as actor non-response (i.e., missing all outgoing ties of an actor), and the the binary covariate was always observed. The six missing data mechanisms are MCAR, MAR related to the covariate, MNAR related to high out-degree, and MNAR related to low out-degree. Further, actors were missing related to high and low in-degree[1].

V. Tested Treatments

In this study, we compare two commonly used naive missing data treatments with multiple imputation.

A. Common Methods

Although the (in)effectiveness of deletion methods has already been explored in multiple studies [8], [11], [13], we incorporate listwise deletion (available cases) in this study, because it is commonly used in network research. It is therefore important to contrast its performance with other methods.

[1]The definitions of MAR and MNAR are problematic in regard to actors missing due to high or low in-degree. The in-degree of all actors is technically unknown when two or more actors are missing, because it is unclear who these actors would have nominated. However, it is still partially observed.

Another commonly used method for handling missing network data is null-tie imputation [13]. In null-tie imputation, all missing links are imputed with zeros. This is comparable to imputing unconditional modes in non-network data, as social networks tend to be sparse with a density below 50%, thus not observing a tie between two actors is the most likely case, ignoring everything else.

B. Multiple Imputation

Multiple imputation is performed using BERGMs following the procedure outlined by Koskinen et al. [15] embedded in the Bergm package in R [6], [20]. In this procedure, the missing network data is imputed using draws from the posterior distribution of the tie variable that is generated to obtain parameter estimates. This procedure was developed for estimation of BERGMs under missing data, however, it is possible to retain the augmented networks, thus achieving proper multiple imputations [21], which contrasts the chosen method from alternatively proposed imputation methods based on ERGMs [22].

For the imputation, we employed two models: A simple model with parameters for density, reciprocity and homophily, and more complex model in which parameters for triadic closure (GWESP) and two-paths (GWDSP) are added to the simple model. In general, multiple imputation should be performed with a model that is at least as complex at the data generating process and contains all parameters that are to be tested in a later step. This ensures that the relationship between the variables is preserved in the imputation. For estimating the imputation models, weakly informative priors $N(0, \sigma = 2)$ were used.

VI. Results

A. Descriptive Outcomes

The performance of the imputation models was inspected for the following descriptive statistics: Average degree, reciprocity (percentage of reciprocated ties), transitivity (percentage of closed two-paths of all two-paths), homophily (percentage of within-group ties on all ties). Further, we evaluate both in-degree and out-degree variance. To measure how the connectivity of the network is preserved by the treatment methods, the average inverse geodesic distance (shortest path from one node to another; both the directed and undirected version) was chosen. Although none of the complete networks has isolated nodes or subgraphs, the inverse geodesic was used because these structures will inevitably appear for higher missing data rates, thus making the shortest path between subgraphs undefined. By taking the inverse these distances will be set to 0. The directed version only follows paths in the direction of the ties, while the undirected version ignores the direction of ties when calculating the geodesics.

Because of spatial limitations we only present the results on an aggregate level combining the results for the 160 networks under each combination missing data mechanism and missing data rate. The pattern of results did not meaningfully differ for the 16 configurations. A detailed analysis of the structural

properties revealed that in most situations, the effect on biases were negligible ($< 5\%$), thus an aggregation was deemed justifiable. The results are presented as average relative bias compared to the statistic calculated on the complete network ($\frac{treated-complete}{complete}$), to obtain comparable results across different network structures[2]. Taking the relative bias is applicable for these statistics because all of them have non-negative scales with 0 as a meaningful endpoint.

All results are presented graphically using grid plots [23], [24]. An example is presented in Fig. 1. Positive values of the relative bias (shown in green) represent an overestimation of the statistics by the treatment and negative values (shown in red) represent an underestimation. For each combination of descriptive statistic and treatment (imputation) method, we created six plots, one for each missing data mechanism, showing the average relative bias for each missing data rate (10-50%). The level of bias is expressed by the saturation of the color, with higher saturation indicating stronger bias. The resulting 48 plots (eight statistics $\times$ six treatment methods) were combined in Fig. 2, presenting all results of the simulation study.

B. Performance of Treatment Methods

The results are presented in Fig. 2. They replicate previous findings (for actor non-response) that measures based on in-degree are more robust than measures based on out-degree, with high degree missing nodes having the strongest effects [3]. Three main conclusions can be drawn: First, for all descriptive statistics, multiple imputation using the complex BERGM performs on average better or equally well compared to any of the other treatment methods. Second, the results indicate that low amounts of missing data (10-20%) can be handled reasonably well by all methods (on average the absolute bias is below 20% over all networks, mechanisms and imputations for all descriptives[3]). Third, homophily was estimated without any relevant bias with all methods under all mechanisms, and is therefore not included in Fig. 2. This does not mean that missing data generally has no effect on measurement of homophily. Specific missing data mechanisms targeting hetero- or homophilous actors or ties can still lead to biased estimates.

Overall, the simple treatments (available cases, null-tie) did not perform well with higher missing data rates. Although in some specific situations they only gave small biases, the nature of missing data mechanisms makes it usually impossible to know which mechanism applies in an empirical setting and thus using these methods is not recommended.

Multiple imputation using the small imputation model performed very well for average degree and reciprocity, but was unable to recover transitivity. The the complex BERGM imputation with the GWESP and GWDSP parameters performed far better on transitivity and other network measures.

[2]We also investigated the average relative absolute bias. Overall, the pattern of results did not change meaningfully.

[3]The exception is the average absolute bias for null-tie imputation for average directed geodesic distance, where the average bias is 22%

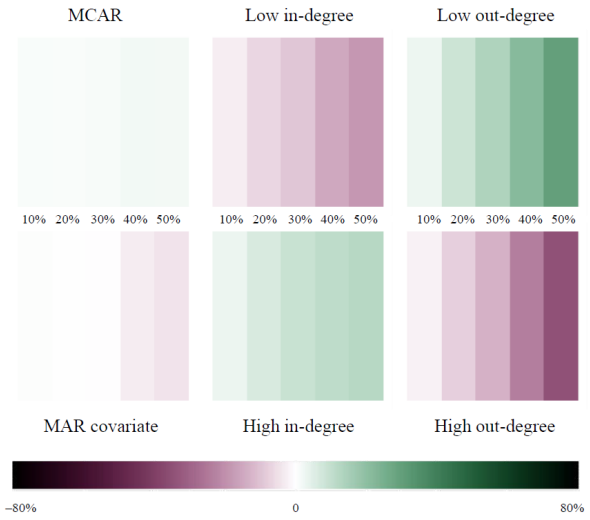

Fig. 1. Example of a grid plot. The six groups consisting of five bars represent the six missing data mechanisms. Each bar stands for 10% missing data, ranging from 10 to 50% missing data. A color scale is given for the interpretation of the colors. The saturation was scaled to 80% as this was the largest average bias observed. Stronger saturation represents larger bias.

VII. Discussion

In this study, we evaluated several missing data treatments for missing network data in a cross-sectional setting. The results indicate that multiple imputation with a sufficiently complex Bayesian ERGM outperforms commonly used techniques. It showed less bias than alternative, naive methods on descriptive statistics, which perform poorly compared to multiple imputation. We expect that the advantage of multiple imputation using Bayesian ERGMs is even stronger in the context of statistical inference, because it is expected to provide more reliable standard errors.

This study focused on a limited number of networks from a restricted set of possible configurations. Moreover, we tested these methods under ideal situations. The biases presented here will probably be larger in empirical data, where the data generating model is not known. Multiple imputation with BERGMs performed especially well, because it represented the data generating model. The results indicated that an insufficiently specified model will not be able to lead to the same reduction in bias than a well-specified imputation model. However, the insufficient model clearly provides more accurate results than the naive methods.

Despite these limitations multiple imputation using BERGMs seems superior to current alternatives in providing unbiased (or less biased) descriptive statistics of networks. The multiple imputation procedure can also be extended to handle missing data in hierarchical [25], longitudinal [26], [27], [28], multilevel [29] and multiplex ERGMs [30]. Future research needs to develop guidelines on the selection of the imputation model and on assessing the sensitivity of the results to model specifications. The results indicate that the imputation with a too simple model will still lead to less bias than the default

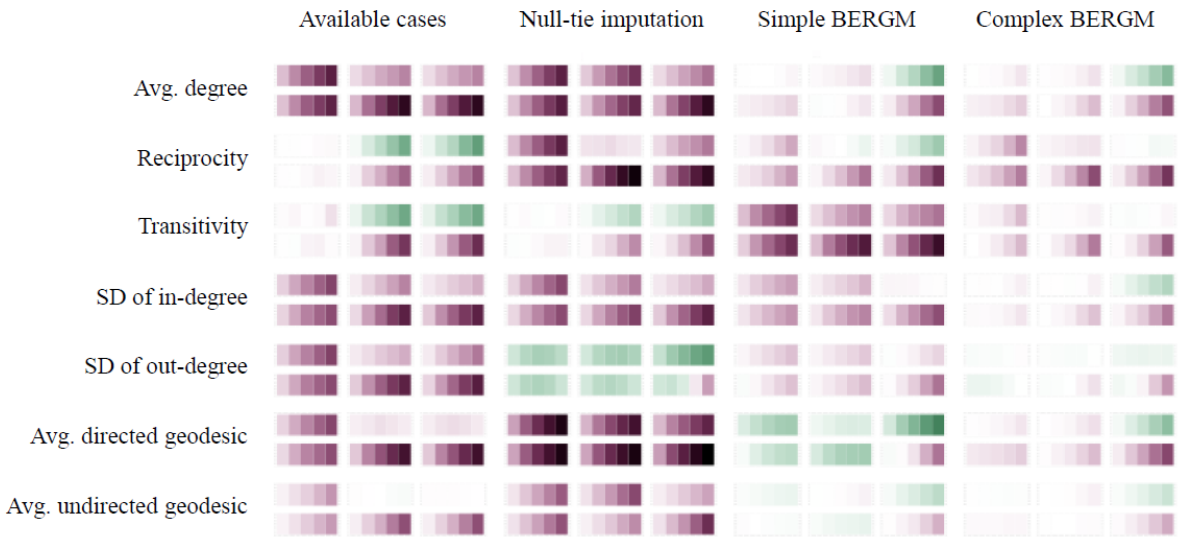

Fig. 2. Average relative bias for each descriptive statistic by treatment method. For interpretation of the plots, consult Fig. 1.

procedures (null-tie imputation and available cases) for the majority of analyzed descriptives.

REFERENCES

[1] M. Huisman and R. W. Krause, "Imputation of missing network data," in Encyclopedia of Social Network Analysis and Mining, R. Alhajj and J. Rokne, Eds. New York: Springer, 2017, pp. 1–10. DOI 10.1007/978-1-4614-7163-9_394-1.

[2] R. W. Krause, M., Huisman, and T. A. B. Snijders, "Multiple Imputation for longitudinal network data," Italian Journal of Applied Statistics, vol. 30, pp. 33–58, 2018.

[3] J. A. Smith, J. Moody, and J. H. Morgan, "Network sampling coverage II: The effect of non-random missing data on network measurement," Social Networks, vol. 48, pp. 78–99, 2017.

[4] D. Lusher, J. H. Koskinen, and G. L. Robins, Exponential random graph models for social networks. New York: Cambridge University Press, 2013.

[5] G. L. Robins, P. Pattison, and J. Woolcock J, "Missing data in networks: exponential random graph ($p*$) models for networks with non-respondents," Social Networks, vol. 26, pp. 257–283, 2004.

[6] A. Caimo and N. Friel, "Bayesian inference for exponential random graph models," Social Networks, vol. 33, pp. 41–55, 2011.

[7] D.B. Rubin, "Inference and missing data," Biometrika, vol. 63, pp. 581–592, 1976.

[8] M. Huisman, and C. E. G. Steglich, "Treatment of non-response in longitudinal network studies," Social Networks, vol. 30, pp. 297–308, 2008.

[9] J. A. Smith and J. Moody, "Structural effects of network sampling coverage I: Nodes missing at random," Social Networks, vol. 35, pp.652–668, 2013.

[10] E. Costenbader and T. W. Valente, "The stability of centrality measures when networks are sampled," Social Networks, vol. 25, pp. 283–307, 2003.

[11] M. Huisman, "Imputation of missing network data: some simple procedures," Journal of Social Structure, vol. 10, pp. 1–29, 2009.

[12] J. L. Schafer and J. W. Graham, "Missing data: our view of the state of the art," Psychological Methods, vol. 7, pp. 147–177, 2002.

[13] A. Žnidaršič, P. Doreian, and A. Ferligoj, "Absent ties in social networks, their treatments, and blockmodeling outcomes," Metodološki Zvezki, vol. 9, pp. 119–138, 2012.

[14] M. S. Handcock and K. Gile, "Modeling social networks from sampled data," Annals of Applied Statistics, vol. 4, pp. 5-25, 2010.

[15] J. H. Koskinen, G. L. Robins, and P. E. Pattison, "Analysing exponential random graph (p-star) models with missing data using Bayesian data augmentation," Statistical Methodology, vol. 7, pp. 366–384, 2010.

[16] S. van Buuren, Flexible Imputation of Missing Data. Boca Raton: Chapman & Hall/CRC, 2012.

[17] M. Handcock, D. Hunter, C. Butts, S. Goodreau, P. Krivitsky, and M. Morris, "Package ergm: Fit, simulate and diagnose exponential-family models for networks", The Statnet Project, 2017. R package version 3.8.0.

[18] R Core Team, "R: A language and environment for statistical computing," R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/, 2018.

[19] T. A. B. Snijders, P. E. Pattison, G. L. Robins, and M. S. Handcock, "New specifications for exponential random graph models," Sociological Methodology, vol. 36, pp. 99–153, 2006.

[20] A. Caimo and N. Friel, "Bergm: Bayesian exponential random graphs in R," Journal of Statistical Software, vol. 61, pp. 1-25, 2014. R package version 4.0.0.

[21] D. B. Rubin, Multiple Imputation for Nonresponse in Surveys. New York: Wiley, 1987.

[22] C. Wang, C. Butts, J. R. Hipp, R. Jose, and C. M. Lakon, "Multiple imputation for missing edge data: a predictive evaluation method with application to add health," Social Networks, vol. 45, pp. 89–98, 2016.

[23] H. Wickham, "ggplot2: Elegant Graphics for Data Analysis," New York: Springer, R package version 3.0.0, 2016.

[24] B. Auguie, "gridExtra: Miscellaneous Functions for 'Grid' Graphics," R package version 2.3, 2017.

[25] M.D. Schweinberger, and M.S. Handcock, "Hierarchical exponential-family random graph models with local dependence," 13, 2012.

[26] A. Caimo, J. Koskinen, and A. Lomi, "Simultaneous inference on network initial conditions and time-dependent dynamics," CASI, pp. 15–16, 2014.

[27] S. Hanneke, W. Fu, and E.P. Xing, "Discrete temporal models of social networks," Electronic Journal of Statistics, vol. 4, pp. 585–605, 2010.

[28] P.N. Krivitsky, and M.S. Handcock, "A separable model for dynamic networks," Journal of the Royal Statistical Society Series B, vol. 76, pp. 29–49, 2013.

[29] P. Wang, G. Robins, P. Pattison, and E. Lazega, "Exponential random graph models for multilevel networks," Social Networks, vol. 35, pp. 96-115, 2013.

[30] P. Wang, "ERGM extensions: models for multiple networks and bipartite networks," in: Exponential Random Graph Models for Social Networks: Theories, Methods and Applications, D. Lusher, J. Koskinen, G.L. Robins, Eds. Cambridge: Cambridge University Press, pp. 115–129, 2013.